

Organização e Recuperação de Informação: Retorno de relevância e expansão de consultas

Marcelo K. A.

Faculdade de Computação - UFU

Relevância: consulta vs. necessidade de informação

- É noção de “relevância para a consulta” é complicada
- **necessidade de informação i** : Busca-se informação sobre se vinho tinto ajudar em reduzir risco de ataque de coração mais que vinho branco.
- **consulta q** : VINHO AND TINTO AND BRANCO AND CORAÇÃO AND ATAQUE
- Considere documento d' : *Ele então entrou no coração de seu discurso: um ataque à indústria por informar sobre como reduzir o risco de beber vinho tinto e branco durante a condução de carros.*
- d' é relevante para a consulta q , mas d' **não** é relevante para a necessidade de informação i .
- Satisfação do usuário somente deve ser avaliada de acordo com a **relevância em relação à necessidade de informação e não à relevância às consultas.**

Veremos hoje

- **Retorno iterativo de relevância:** melhorar resultados iniciais por meio da indicação ao sistema de ORI de quais documentos são relevantes ou não

Veremos hoje

- **Retorno iterativo de relevância:** melhorar resultados iniciais por meio da indicação ao sistema de ORI de quais documentos são relevantes ou não

Veremos hoje

- **Retorno iterativo de relevância:** melhorar resultados iniciais por meio da indicação ao sistema de ORI de quais documentos são relevantes ou não
- Método mais conhecido: **Rocchio feedback**

Veremos hoje

- **Retorno iterativo de relevância:** melhorar resultados iniciais por meio da indicação ao sistema de ORI de quais documentos são relevantes ou não
- Método mais conhecido: **Rocchio feedback**
- **Expansão de consulta:** melhorar recuperação resultados ao adicionar sinônimos / relacionados termos à consulta

Veremos hoje

- **Retorno iterativo de relevância:** melhorar resultados iniciais por meio da indicação ao sistema de ORI de quais documentos são relevantes ou não
- Método mais conhecido: **Rocchio feedback**
- **Expansão de consulta:** melhorar recuperação resultados ao adicionar sinônimos / relacionados termos à consulta
 - **Fontes para termos relacionados:** tesouros manuais, tesouros automáticos, históricos de consultas

Sumário

- 1 Motivação
- 2 Retorno de relevância: básico
- 3 Retorno de relevância: detalhes
- 4 Expansão de consulta

Como melhorar a taxa de recuperação na busca?

- Tópico principal de hoje: duas formas para melhorar a recuperação: usar retorno de relevância obtido com usuários e usar expansão de consulta

Como melhorar a taxa de recuperação na busca?

- Tópico principal de hoje: duas formas para melhorar a recuperação: usar retorno de relevância obtido com usuários e usar expansão de consulta
- Como exemplo considere uma consulta q : [aeronave] ...

Como melhorar a taxa de recuperação na busca?

- Tópico principal de hoje: duas formas para melhorar a recuperação: usar retorno de relevância obtido com usuários e usar expansão de consulta
- Como exemplo considere uma consulta q : [aeronave] ...
- ... e um documento d contendo “avião”, mas não contendo “aeronave”

Como melhorar a taxa de recuperação na busca?

- Tópico principal de hoje: duas formas para melhorar a recuperação: usar retorno de relevância obtido com usuários e usar expansão de consulta
- Como exemplo considere uma consulta q : [aeronave] ...
- ... e um documento d contendo “avião”, mas não contendo “aeronave”
- Um sistema de ORI simples não retorna d para q .

Como melhorar a taxa de recuperação na busca?

- Tópico principal de hoje: duas formas para melhorar a recuperação: usar retorno de relevância obtido com usuários e usar expansão de consulta
- Como exemplo considere uma consulta q : [aeronave] ...
- ... e um documento d contendo “avião”, mas não contendo “aeronave”
- Um sistema de ORI simples não retorna d para q .
- Embora d seja o documento mais relevante para q !

Como melhorar a taxa de recuperação na busca?

- Tópico principal de hoje: duas formas para melhorar a recuperação: usar retorno de relevância obtido com usuários e usar expansão de consulta
- Como exemplo considere uma consulta q : [aeronave] ...
- ... e um documento d contendo “avião”, mas não contendo “aeronave”
- Um sistema de ORI simples não retorna d para q .
- Embora d seja o documento mais relevante para q !
- Queremos mudar isso

Como melhorar a taxa de recuperação na busca?

- Tópico principal de hoje: duas formas para melhorar a recuperação: usar retorno de relevância obtido com usuários e usar expansão de consulta
- Como exemplo considere uma consulta q : [aeronave] ...
- ... e um documento d contendo “avião”, mas não contendo “aeronave”
- Um sistema de ORI simples não retorna d para q .
- Embora d seja o documento mais relevante para q !
- Queremos mudar isso
 - Retornar documentos relevantes mesmo se não houver termos iguais aos da consulta original

Taxa de recuperação

- Queremos: “aumentar o número de documentos relevantes retornados para o usuário”

Taxa de recuperação

- Queremos: “aumentar o número de documentos relevantes retornados para o usuário”
- por exemplo, expandir a consulta “jaguar” para incluir “pantera”

Taxa de recuperação

- Queremos: “aumentar o número de documentos relevantes retornados para o usuário”
- por exemplo, expandir a consulta “jaguar” para incluir “pantera”
 - . . . isso pode eliminar alguns documentos relevantes, mas em geral aumenta os documentos relevantes retornados nas páginas “top”

Opções para melhorar recuperação

- Opção local: fazer uma análise “local”, sob-demanda para a consulta

Opções para melhorar recuperação

- Opção local: fazer uma análise “local”, sob-demanda para a consulta
 - Método local principal: obter retorno (opinião do usuário) sobre a relevância

Opções para melhorar recuperação

- Opção local: fazer uma análise “local”, sob-demanda para a consulta
 - Método local principal: obter retorno (opinião do usuário) sobre a relevância
 - Parte 1

Opções para melhorar recuperação

- Opção local: fazer uma análise “local”, sob-demanda para a consulta
 - Método local principal: obter retorno (opinião do usuário) sobre a relevância
 - Parte 1
- Global: fazer uma análise global (por exemplo, da coleção inteira) para produzir um **tesauro**

Opções para melhorar recuperação

- Opção local: fazer uma análise “local”, sob-demanda para a consulta
 - Método local principal: obter retorno (opinião do usuário) sobre a relevância
 - Parte 1
- Global: fazer uma análise global (por exemplo, da coleção inteira) para produzir um **tesauro**
- **Tesauro** é um dicionário/lista termos com ideias e significados semelhantes

Opções para melhorar recuperação

- Opção local: fazer uma análise “local”, sob-demanda para a consulta
 - Método local principal: obter retorno (opinião do usuário) sobre a relevância
 - Parte 1
- Global: fazer uma análise global (por exemplo, da coleção inteira) para produzir um tesouro
- Tesouro é um dicionário/lista termos com ideias e significados semelhantes
 - Usar um tesouro para expansão de consulta

Sumário

- 1 Motivação
- 2 Retorno de relevância: básico
- 3 Retorno de relevância: detalhes
- 4 Expansão de consulta

Retorno de relevância: ideia básica

- 1 O usuário faz um consulta.

Retorno de relevância: ideia básica

- 1 O usuário faz um consulta.
- 2 O buscador retorna um conjunto de documentos.

Retorno de relevância: ideia básica

- 1 O usuário faz um consulta.
- 2 O buscador retorna um conjunto de documentos.
- 3 Usuário marca alguns documentos como relevantes e outros como irrelevantes.

Retorno de relevância: ideia básica

- 1 O usuário faz um consulta.
- 2 O buscador retorna um conjunto de documentos.
- 3 Usuário marca alguns documentos como relevantes e outros como irrelevantes.
- 4 Buscador computa nova representação da consulta.
Esperança: que seja melhor que a consulta inicial.

Retorno de relevância: ideia básica

- 1 O usuário faz um consulta.
- 2 O buscador retorna um conjunto de documentos.
- 3 Usuário marca alguns documentos como relevantes e outros como irrelevantes.
- 4 Buscador computa nova representação da consulta.
Esperança: que seja melhor que a consulta inicial.
- 5 Buscador retorna novos resultados para a nova consulta.

Retorno de relevância: ideia básica

- 1 O usuário faz um consulta.
- 2 O buscador retorna um conjunto de documentos.
- 3 Usuário marca alguns documentos como relevantes e outros como irrelevantes.
- 4 Buscador computa nova representação da consulta.
Esperança: que seja melhor que a consulta inicial.
- 5 Buscador retorna novos resultados para a nova consulta.
- 6 Novos resultados têm (se tudo funcionar bem) melhor taxa de recuperação.

Retorno de relevância

- Podemos repetir esse procedimento: vários turnos de retorno de relevância.

Retorno de relevância

- Podemos repetir esse procedimento: vários turnos de retorno de relevância.
- Usaremos o termo **recuperação ad hoc** para referir à recuperação sem a obtenção de retorno de relevância.

Retorno de relevância













- Podemos repetir esse procedimento: vários turnos de retorno de relevância.
- Usaremos o termo **recuperação ad hoc** para referir à recuperação sem a obtenção de retorno de relevância.
- Veremos exemplos de retorno de relevância que destacam diferentes aspectos do processo.

Retorno de relevância: Exemplo 1















Resultados para consulta inicial

[Browse](#)
[Search](#)
[Prev](#)
[Next](#)
[Random](#)













					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0

Retorno do usuário: selecionar o que é relevante

[Browse](#)
[Search](#)
[Prev](#)
[Next](#)
[Random](#)

					
(144473, 16458)	(144457, 252140)	(144456, 262857)	(144456, 262863)	(144457, 252134)	(144483, 265154)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
					
(144483, 264644)	(144483, 265153)	(144518, 257752)	(144538, 525937)	(144456, 249611)	(144456, 250064)
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0

Resultados após retorno de relevância

Browse Search Prev Next Random					
					
(144538, 523493) 0.54182 0.231944 0.309876	(144538, 523835) 0.56319296 0.267304 0.295889	(144538, 523529) 0.584279 0.280881 0.303398	(144456, 253569) 0.64501 0.351395 0.293615	(144456, 253568) 0.650275 0.411745 0.23853	(144538, 523799) 0.66709197 0.358033 0.309059
					
(144473, 16249) 0.6721 0.393922 0.278178	(144456, 249634) 0.675018 0.4659 0.211118	(144456, 253693) 0.676901 0.47645 0.200451	(144473, 16328) 0.700339 0.309002 0.391357	(144483, 265264) 0.70170796 0.36176 0.339948	(144478, 512410) 0.70297 0.469111 0.233859

Um exemplo textual

Consulta inicial: [new space satellite applications]

Um exemplo textual

Consulta inicial: [new space satellite applications]

Um exemplo textual

Consulta inicial: [new space satellite applications]

Resultados para consulta inicial: ($r = \text{rank}$)

r

- | | | |
|---|-------|--|
| 1 | 0.539 | NASA Hasn't Scrapped Imaging Spectrometer |
| 2 | 0.533 | NASA Scratches Environment Gear from Satellite Plan |
| 3 | 0.528 | Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes |
| 4 | 0.526 | A NASA Satellite Project Accomplishes Incredible Feat: Staying within Budget |
| 5 | 0.525 | Scientist Who Exposed Global Warming Proposes Satellites for Climate Research |
| 6 | 0.524 | Report Provides Support for the Critics Of Using Big Satellites to Study Climate |
| 7 | 0.516 | Arianespace Receives Satellite Launch Pact from Telesat Canada |
| 8 | 0.509 | Telecommunications Tale of two Companies |

Um exemplo textual

Consulta inicial: [new space satellite applications]

Resultados para consulta inicial: (r = rank)

r		
1	0.539	NASA Hasn't Scrapped Imaging Spectrometer
2	0.533	NASA Scratches Environment Gear from Satellite Plan
3	0.528	Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4	0.526	A NASA Satellite Project Accomplishes Incredible Feat: Staying within Budget
5	0.525	Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
6	0.524	Report Provides Support for the Critics Of Using Big Satellites to Study Climate
7	0.516	Arianespace Receives Satellite Launch Pact from Telesat Canada
8	0.509	Telecommunications Tale of two Companies

Usuário então marca documentos relevantes com “+”.

Um exemplo textual

Consulta inicial: [new space satellite applications]

Resultados para consulta inicial: (r = rank)

	r		
+	1	0.539	NASA Hasn't Scrapped Imaging Spectrometer
+	2	0.533	NASA Scratches Environment Gear from Satellite Plan
	3	0.528	Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
	4	0.526	A NASA Satellite Project Accomplishes Incredible Feat: Staying within Budget
	5	0.525	Scientist Who Exposed Global Warming Proposes Satellites for Climate Research
	6	0.524	Report Provides Support for the Critics Of Using Big Satellites to Study Climate
	7	0.516	Arianespace Receives Satellite Launch Pact from Telesat Canada
+	8	0.509	Telecommunications Tale of two Companies

Usuário então marca documentos relevantes com “+”.

Consulta expandida após o retorno de relevância

2.074	new	15.106	space
30.816	satellite	5.660	application
5.991	nasa	5.196	eos
4.196	launch	3.972	aster
3.516	instrument	3.446	arianespace
3.004	bundespost	2.806	ss
2.790	rocket	2.053	scientist
2.003	broadcast	1.172	earth
0.836	oil	0.646	measure

Comparar à consulta original: [new space satellite applications]

Resultados para consulta expandida (posição anterior entre parêntesis)

	<i>r</i>		
*	1 (2)	0.513	NASA Scratches Environment Gear de Satellite Plan
*	2 (1)	0.500	NASA Hasn't Scrapped Imaging Spectrometer
	3	0.493	When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
	4	0.493	NASA Uses 'Warm' Superconductors for Fast Circuit
*	5 (8)	0.492	Telecommunications Tale of two Companies
	6	0.491	Soviets May Adapt Parts of SS-20 Missile For Commercial Use
	7	0.490	Gaping Gap: Pentagon Lags in Race To Match the Soviets in Rocket Launchers
	8	0.490	Rescue of Satellite By Space Agency To Cost \$90

Sumário

- 1 Motivação
- 2 Retorno de relevância: básico
- 3 Retorno de relevância: detalhes
- 4 Expansão de consulta

Conceito-chave para retorno de relevância: centroide

- O centroide é o centro de massa de um conjunto de vetores.

Conceito-chave para retorno de relevância: centroide

- O centroide é o centro de massa de um conjunto de vetores.
- Recuperar documentos representados como vetores em um espaço de alta-dimensionalidade

Conceito-chave para retorno de relevância: centroide

- O centroide é o centro de massa de um conjunto de vetores.
- Recuperar documentos representados como vetores em um espaço de alta-dimensionalidade
- Então: computamos centroides de documentos.

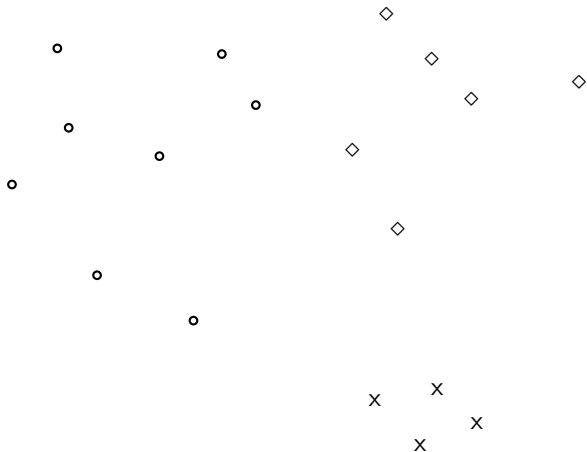
Conceito-chave para retorno de relevância: centroide

- O centroide é o centro de massa de um conjunto de vetores.
- Recuperar documentos representados como vetores em um espaço de alta-dimensionalidade
- Então: computamos centroides de documentos.
- Definição:

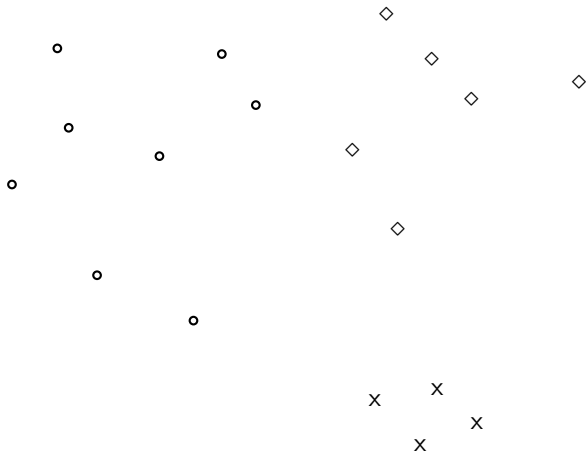
$$\vec{\mu}(D) = \frac{1}{|D|} \sum_{d \in D} \vec{v}(d)$$

onde D é um conjunto de documentos e $\vec{v}(d) = \vec{d}$ é o vetor que usamos para representar um documento d .

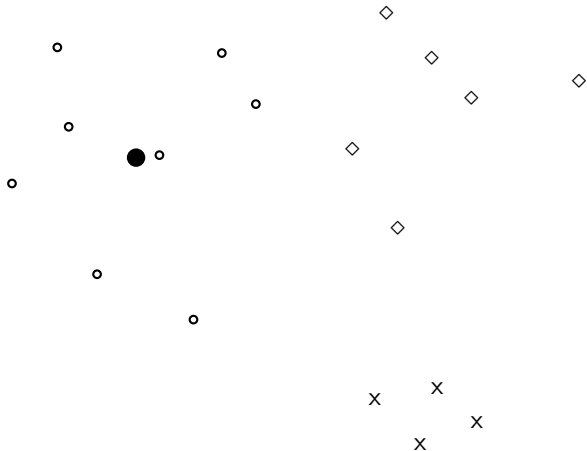
Centroide: Exemplos



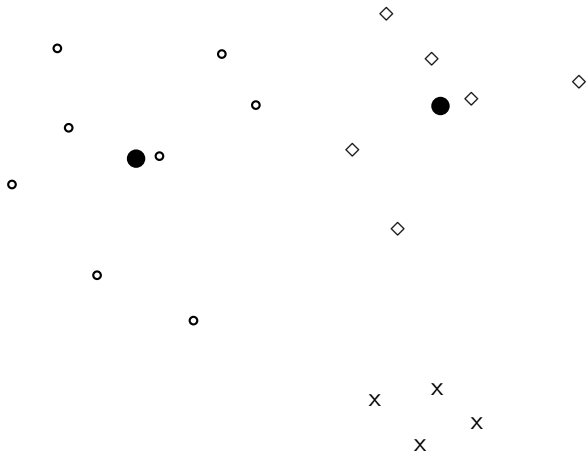
Centroide: Exemplos



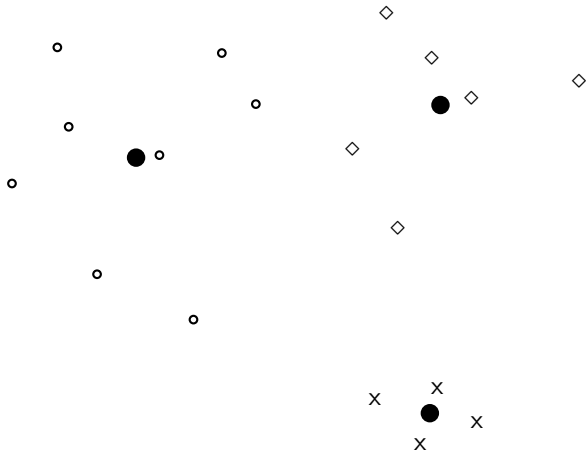
Centroide: Exemplos



Centroide: Exemplos



Centroide: Exemplos



Algoritmo de Rocchio

- O Algoritmo de Rocchio implementa o retorno de relevância no modelo vetorial.

Algoritmo de Rocchio

- O Algoritmo de Rocchio implementa o retorno de relevância no modelo vetorial.
- Rocchio procura pela consulta \vec{q}_{opt} que maximiza

$$\vec{q}_{opt} = \arg \max_{\vec{q}} [\text{sim}(\vec{q}, \mu(D_r)) - \text{sim}(\vec{q}, \mu(D_{nr}))]$$

D_r : conjunto de documentos relevantes; D_{nr} : conjunto de documentos não relevantes; $\text{sim}()$ é uma função de similaridade

Algoritmo de Rocchio

- O Algoritmo de Rocchio implementa o retorno de relevância no modelo vetorial.
- Rocchio procura pela consulta \vec{q}_{opt} que maximiza

$$\vec{q}_{opt} = \arg \max_{\vec{q}} [\text{sim}(\vec{q}, \mu(D_r)) - \text{sim}(\vec{q}, \mu(D_{nr}))]$$

D_r : conjunto de documentos relevantes; D_{nr} : conjunto de documentos não relevantes; $\text{sim}()$ é uma função de similaridade

- Intenção: \vec{q}_{opt} é o vetor que melhor separa documentos relevantes de não relevantes

Algoritmo de Rocchio

- O Algoritmo de Rocchio implementa o retorno de relevância no modelo vetorial.
- Rocchio procura pela consulta \vec{q}_{opt} que maximiza

$$\vec{q}_{opt} = \arg \max_{\vec{q}} [\text{sim}(\vec{q}, \mu(D_r)) - \text{sim}(\vec{q}, \mu(D_{nr}))]$$

D_r : conjunto de documentos relevantes; D_{nr} : conjunto de documentos não relevantes; $\text{sim}()$ é uma função de similaridade

- Intenção: \vec{q}_{opt} é o vetor que melhor separa documentos relevantes de não relevantes
- Assumindo outras premissas, podemos reescrever \vec{q}_{opt} como:

$$\vec{q}_{opt} = \mu(D_r) + [\mu(D_r) - \mu(D_{nr})]$$

Algoritmo de Rocchio

- O vetor ótimo de consulta é:

$$\begin{aligned}\vec{q}_{opt} &= \mu(D_r) + [\mu(D_r) - \mu(D_{nr})] \\ &= \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + \left[\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j \right]\end{aligned}$$

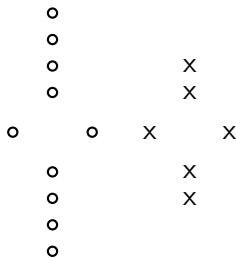
Algoritmo de Rocchio

- O vetor ótimo de consulta é:

$$\begin{aligned}\vec{q}_{opt} &= \mu(D_r) + [\mu(D_r) - \mu(D_{nr})] \\ &= \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j + \left[\frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j \right]\end{aligned}$$

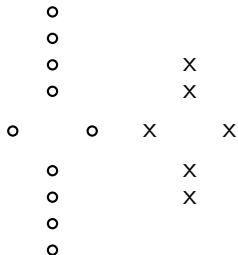
- Movemos o centroide dos documentos relevantes pela diferença entre os dois centroides.

Exercício: Calcular vetor de Rocchio



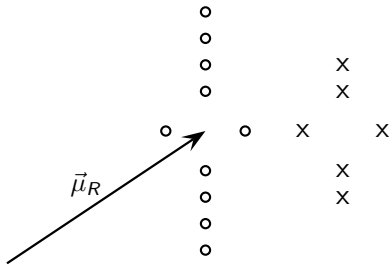
Círculos: documentos relevantes, Xs: documentos não relevantes

Rocchio



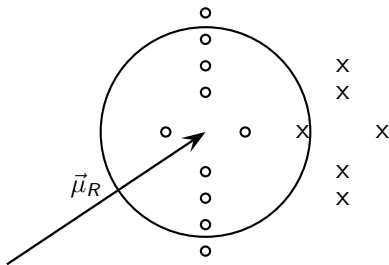
círculos: documentos relevantes, Xs: documentos não relevantes

Rocchio



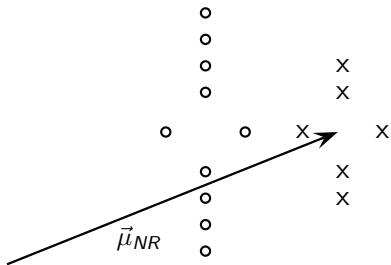
$\vec{\mu}_R$: centroide de documentos relevantes

Rocchio



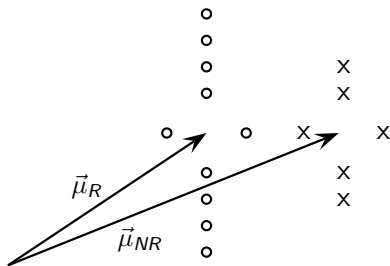
$\vec{\mu}_R$ não separa relevantes/não-relevantes.

Rocchio

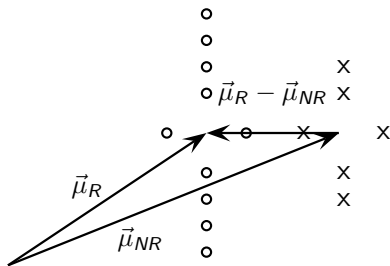


$\vec{\mu}_{NR}$: centroide de documentos não relevantes

Rocchio

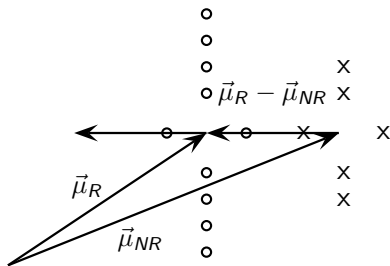


Rocchio



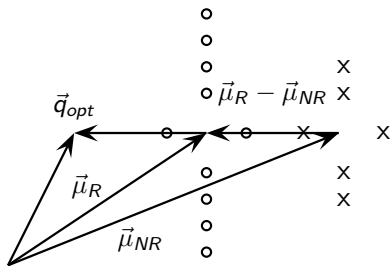
$\vec{\mu}_R - \vec{\mu}_{NR}$: vetor-diferença

Rocchio



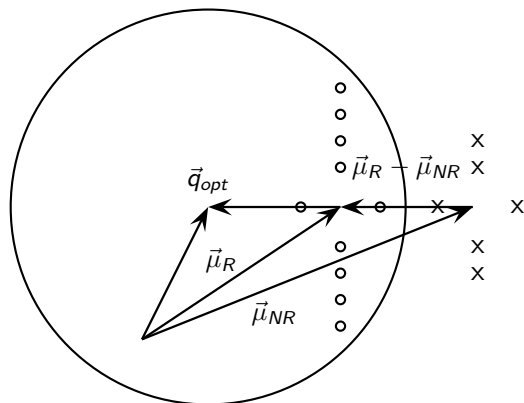
Somar o vetor-diferença em $\vec{\mu}_R \dots$

Rocchio



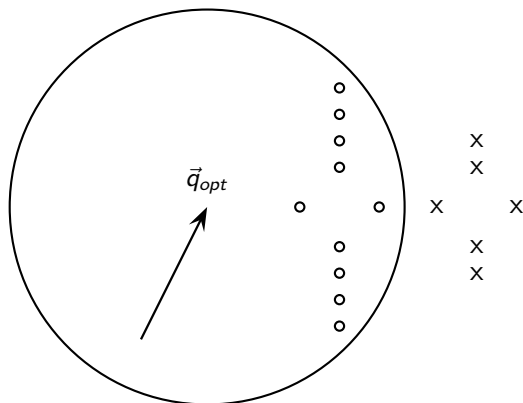
... para obter \vec{q}_{opt}

Rocchio



\vec{q}_{opt} separa relevantes/não-relevantes perfeitamente.

Rocchio



\vec{q}_{opt} separa relevantes/não-relevantes perfeitamente.

Rocchio 1971 algoritmo (versão SMART)

- Usado na prática:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : vetor de consulta modificado; q_0 : vetor de consulta original ; D_r e D_{nr} : conjuntos de documentos conhecidos como relevantes e não relevantes, respectivamente; α , β , e γ : pesos

Rocchio 1971 algoritmo (versão SMART)

- Usado na prática:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : vetor de consulta modificado; q_0 : vetor de consulta original ; D_r e D_{nr} : conjuntos de documentos conhecidos como relevantes e não relevantes, respectivamente; α , β , e γ : pesos

- nova consulta move na direção de documentos relevantes e longe de documentos irrelevantes.

Rocchio 1971 algoritmo (versão SMART)

- Usado na prática:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : vetor de consulta modificado; q_0 : vetor de consulta original ; D_r e D_{nr} : conjuntos de documentos conhecidos como relevantes e não relevantes, respectivamente; α , β , e γ : pesos

- nova consulta move na direção de documentos relevantes e longe de documentos irrelevantes.
- Equilíbrio entre α vs. β/γ : se temos muitos documentos avaliados, precisamos de maior relação β/γ .

Rocchio 1971 algoritmo (versão SMART)

- Usado na prática:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : vetor de consulta modificado; q_0 : vetor de consulta original ; D_r e D_{nr} : conjuntos de documentos conhecidos como relevantes e não relevantes, respectivamente; α , β , e γ : pesos

- nova consulta move na direção de documentos relevantes e longe de documentos irrelevantes.
- Equilíbrio entre α vs. β/γ : se temos muitos documentos avaliados, precisamos de maior relação β/γ .
- Atribuir 0 para os pesos negativos.

Rocchio 1971 algoritmo (versão SMART)

- Usado na prática:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : vetor de consulta modificado; q_0 : vetor de consulta original ; D_r e D_{nr} : conjuntos de documentos conhecidos como relevantes e não relevantes, respectivamente; α , β , e γ : pesos

- nova consulta move na direção de documentos relevantes e longe de documentos irrelevantes.
- Equilíbrio entre α vs. β/γ : se temos muitos documentos avaliados, precisamos de maior relação β/γ .
- Atribuir 0 para os pesos negativos.
- “Peso negativo” não faz sentido no modelo vetorial.

Rocchio 1971 algoritmo (versão SMART)

- Usado na prática:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : vetor de consulta modificado; q_0 : vetor de consulta original ; D_r e D_{nr} : conjuntos de documentos conhecidos como relevantes e não relevantes, respectivamente; α , β , e γ : pesos

- nova consulta move na direção de documentos relevantes e longe de documentos irrelevantes.
- Equilíbrio entre α vs. β/γ : se temos muitos documentos avaliados, precisamos de maior relação β/γ .
- Atribuir 0 para os pesos negativos.
- “Peso negativo” não faz sentido no modelo vetorial.

Rocchio 1971 algoritmo (versão SMART)

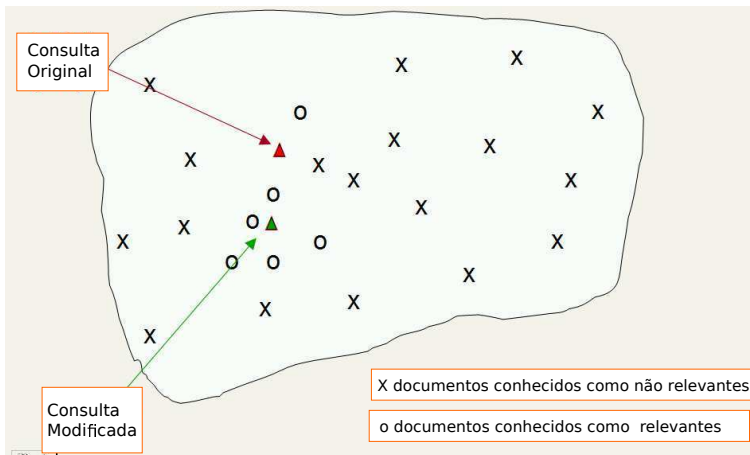
- Usado na prática:

$$\begin{aligned}\vec{q}_m &= \alpha \vec{q}_0 + \beta \mu(D_r) - \gamma \mu(D_{nr}) \\ &= \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j\end{aligned}$$

q_m : vetor de consulta modificado; q_0 : vetor de consulta original ; D_r e D_{nr} : conjuntos de documentos conhecidos como relevantes e não relevantes, respectivamente; α , β , e γ : pesos

- nova consulta move na direção de documentos relevantes e longe de documentos irrelevantes.
- Equilíbrio entre α vs. β/γ : se temos muitos documentos avaliados, precisamos de maior relação β/γ .
- **Atribuir 0 para os pesos negativos.**
- “Peso negativo” não faz sentido no modelo vetorial.

Rocchio



Retorno de relevância positivo vs. negativo

- Retorno positivo tem maior valor que o retorno negativo.

Retorno de relevância positivo vs. negativo

- Retorno positivo tem maior valor que o retorno negativo.
- Por exemplo, configurar $\beta = 0.75$, $\gamma = 0.25$ para dar maior peso retorno positivo. Estratégia de reforço positivo.

Retorno de relevância positivo vs. negativo

- Retorno positivo tem maior valor que o retorno negativo.
- Por exemplo, configurar $\beta = 0.75$, $\gamma = 0.25$ para dar maior peso retorno positivo. Estratégia de reforço positivo.
- Muitos sistemas permitem somente retorno positivo.

Retorno de Relevância: problemas

- Retorno de relevância é caro.

Retorno de Relevância: problemas

- Retorno de relevância é caro.
 - Retorno de relevância faz com que consultas tornem-se longas (porquê?)

Retorno de Relevância: problemas

- Retorno de relevância é caro.
 - Retorno de relevância faz com que consultas tornem-se longas (porquê?)
 - Consultas longas são caras para processar.

Retorno de Relevância: problemas

- Retorno de relevância é caro.
 - Retorno de relevância faz com que consultas tornem-se longas (porquê?)
 - Consultas longas são caras para processar.
- Usuários evitam fornecer retorno explícito.

Retorno de Relevância: problemas

- Retorno de relevância é caro.
 - Retorno de relevância faz com que consultas tornem-se longas (porquê?)
 - Consultas longas são caras para processar.
- Usuários evitam fornecer retorno explícito.
- Frequentemente fica difícil entender porque um determinado documento foi retornado após aplicar a informação obtida com o retorno de relevância

Pseudo-retorno de relevância

- O retorno de pseudo-relevância automatiza a parte “manual” do retorno da verdadeira relevância.

Pseudo-retorno de relevância

- O retorno de pseudo-relevância automatiza a parte “manual” do retorno da verdadeira relevância.
- Algoritmo de pseudo-relevância:

Pseudo-retorno de relevância

- O retorno de pseudo-relevância automatiza a parte “manual” do retorno da verdadeira relevância.
- Algoritmo de pseudo-relevância:
 - Recupera a lista ordenação de resultados para a consulta do usuário

Pseudo-retorno de relevância

- O retorno de pseudo-relevância automatiza a parte “manual” do retorno da verdadeira relevância.
- Algoritmo de pseudo-relevância:
 - Recupera a lista ordenação de resultados para a consulta do usuário
 - Assume que os top k documentos são relevantes.

Pseudo-retorno de relevância

- O retorno de pseudo-relevância automatiza a parte “manual” do retorno da verdadeira relevância.
- Algoritmo de pseudo-relevância:
 - Recupera a lista ordenação de resultados para a consulta do usuário
 - Assume que os top k documentos são relevantes.
 - Retorno de relevância (e.g., Rocchio)

Pseudo-retorno de relevância

- O retorno de pseudo-relevância automatiza a parte “manual” do retorno da verdadeira relevância.
- Algoritmo de pseudo-relevância:
 - Recupera a lista ordenação de resultados para a consulta do usuário
 - Assume que os top k documentos são relevantes.
 - Retorno de relevância (e.g., Rocchio)
- Funciona bem em média, mas pode falhar para algumas consultas.

Pseudo-retorno de relevância

- O retorno de pseudo-relevância automatiza a parte “manual” do retorno da verdadeira relevância.
- Algoritmo de pseudo-relevância:
 - Recupera a lista ordenação de resultados para a consulta do usuário
 - Assume que os top k documentos são relevantes.
 - Retorno de relevância (e.g., Rocchio)
- Funciona bem em média, mas pode falhar para algumas consultas.
- Várias iterações pode causar um *desvio de consulta*.

Sumário

- 1 Motivação
- 2 Retorno de relevância: básico
- 3 Retorno de relevância: detalhes
- 4 Expansão de consulta

Expansão de consulta

- Expansão de consulta é outro método para **aumentar a recuperação**.

Expansão de consulta

- Expansão de consulta é outro método para **aumentar a recuperação**.
- Usamos “expansão global de consultas” para referir a “métodos globais de reformulação de consulta”.

Expansão de consulta

- Expansão de consulta é outro método para **aumentar a recuperação**.
- Usamos “expansão global de consultas” para referir a “métodos globais de reformulação de consulta”.
- Em expansão global de consultas, uma consulta é modificada baseada em alguma característica global da coleção, isto é, um recurso que não é não dependente da consulta.

Expansão de consulta

- Expansão de consulta é outro método para **aumentar a recuperação**.
- Usamos “expansão global de consultas” para referir a “métodos globais de reformulação de consulta”.
- Em expansão global de consultas, uma consulta é modificada baseada em alguma característica global da coleção, isto é, um recurso que não é não dependente da consulta.
- Informação principal que usamos: (quase-)sinônimo

Expansão de consulta

- Expansão de consulta é outro método para **aumentar a recuperação**.
- Usamos “expansão global de consultas” para referir a “métodos globais de reformulação de consulta”.
- Em expansão global de consultas, uma consulta é modificada baseada em alguma característica global da coleção, isto é, um recurso que não é não dependente da consulta.
- Informação principal que usamos: (quase-)sinônimo
- Uma publicação ou base de dados que lista (quase-)sinônimos é um **tesauro**.

Expansão de consulta

- Expansão de consulta é outro método para **aumentar a recuperação**.
- Usamos “expansão global de consultas” para referir a “métodos globais de reformulação de consulta”.
- Em expansão global de consultas, uma consulta é modificada baseada em alguma característica global da coleção, isto é, um recurso que não é não dependente da consulta.
- Informação principal que usamos: (quase-)sinônimo
- Uma publicação ou base de dados que lista (quase-)sinônimos é um **tesauro**.
- Olharemos dois tipos de tesouros: 1) criados manualmente e 2) automaticamente.

Expansão de consulta: exemplo

YAHOO! SEARCH

Web | [Images](#) | [Video](#) | [Audio](#) | [Directory](#) | [Local](#) | [News](#) | [Shopping](#) | [More »](#)

[Answers](#) | [My Web](#) | [Search Services](#) | [Advanced Search](#) | [Preferences](#)

Search Results 1 - 10 of about 160,000,000 for [palm](#) - 0.07 sec. ([About this page](#))

Also try: [palm springs](#), [palm pilot](#), [palm trees](#), [palm reading](#) [More...](#)

SPONSOR RESULTS

- [Official Palm Store](#)
[store.palm.com](#) Free shipping on all handhelds and more at the official **Palm** store.
- [Palms Hotel - Best Rate Guarantee](#)
[www.vegas.com](#) Book the **Palms** Hotel Casino with our best rate guarantee at VEGAS.com, the official Vegas travel site.

 [Palm Pilots](#) - [Palm Downloads](#)
[Yahoo! Shortcut](#) - [About](#)

- [Palm, Inc.](#) [Ⓜ]
Maker of handheld PDA devices that allow mobile users to manage schedules, contacts, and other personal and business information.
Category: [B2B](#) > [Personal Digital Assistants \(PDAs\)](#)
[www.palm.com](#) - 20k - [Cached](#) - [More from this site](#) - [Save](#)

SPONSOR RESULTS

[Palm Memory](#)
Memory Giant is fast and easy. Guaranteed compatible memory. Great...
[www.memorygiant.com](#)

[The Palms, Turks and Caicos Islands](#)
Resort/Condo photos, rates, availability and reservations....
[www.worldwidereservationsystems.c](#)

[The Palms Casino Resort, Las Vegas](#)
Low price guarantee at the **Palms** Casino resort in Las Vegas. Book...
[lasvegas.hotelscorp.com](#)

Tipos de retorno de usuário

- Usuário dá retorno sobre [documentos](#).

Tipos de retorno de usuário

- Usuário dá retorno sobre [documentos](#).
 - Mais comum em retorno de relevância

Tipos de retorno de usuário

- Usuário dá retorno sobre **documentos**.
 - Mais comum em retorno de relevância
- Usuário dá retorno sobre **palavras** ou **expressões**.

Tipos de retorno de usuário

- Usuário dá retorno sobre **documentos**.
 - Mais comum em retorno de relevância
- Usuário dá retorno sobre **palavras** ou **expressões**.
 - Mais comum em expansão de consulta

Tipos de expansão de consulta

- Tesouro construído manualmente (mantido por editores, e.g., PubMed)

Tipos de expansão de consulta

- Tesouro construído manualmente (mantido por editores, e.g., PubMed)
- Tesouro construído automaticamente (e.g., baseado em estatísticas de co-ocorrência)

Tipos de expansão de consulta

- Tesouro construído manualmente (mantido por editores, e.g., PubMed)
- Tesouro construído automaticamente (e.g., baseado em estatísticas de co-ocorrência)
- Equivalência de consulta baseada em mineração de histórico de consultas (comum na web como no exemplo “palm”)

Expansão de consulta com tesouros

- Para cada termo t na consulta, expandir a consulta com palavras relacionados com t nas listas de tesouro.

Expansão de consulta com tesouros

- Para cada termo t na consulta, expandir a consulta com palavras relacionados com t nas listas de tesouro.
- HOSPITAL \rightarrow MÉDICO

Expansão de consulta com tesouros

- Para cada termo t na consulta, expandir a consulta com palavras relacionados com t nas listas de tesouro.
- HOSPITAL \rightarrow MÉDICO
- Em geral aumenta recuperação

Expansão de consulta com tesouros

- Para cada termo t na consulta, expandir a consulta com palavras relacionados com t nas listas de tesouro.
- HOSPITAL \rightarrow MÉDICO
- Em geral aumenta recuperação
- Pode diminuir precisão, especificamente com ambíguos termos

Expansão de consulta com tesouros

- Para cada termo t na consulta, expandir a consulta com palavras relacionados com t nas listas de tesouro.
- HOSPITAL → MÉDICO
- Em geral aumenta recuperação
- Pode diminuir precisão, especificamente com ambíguos termos
 - INTEREST RATE → INTEREST RATE FASCINATE

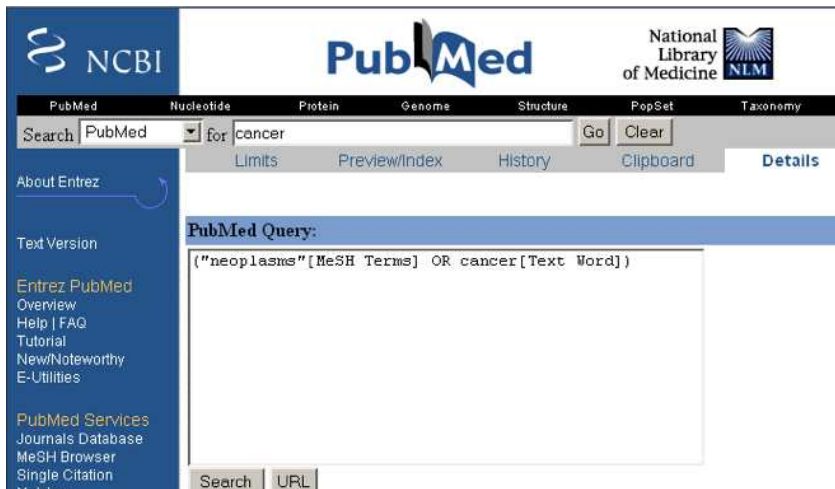
Expansão de consulta com tesouros

- Para cada termo t na consulta, expandir a consulta com palavras relacionados com t nas listas de tesouro.
- HOSPITAL → MÉDICO
- Em geral aumenta recuperação
- Pode diminuir precisão, especificamente com ambíguos termos
 - INTEREST RATE → INTEREST RATE FASCINATE
- Amplamente utilizado em motores de busca especializados para ciência e engenharia

Expansão de consulta com tesouros

- Para cada termo t na consulta, expandir a consulta com palavras relacionados com t nas listas de tesouro.
- HOSPITAL → MÉDICO
- Em geral aumenta recuperação
- Pode diminuir precisão, especificamente com ambíguos termos
 - INTEREST RATE → INTEREST RATE FASCINATE
- Amplamente utilizado em motores de busca especializados para ciência e engenharia
- É muito caro criar e manter um tesouro manualmente

Exemplo de tesouro manual: PubMed



The screenshot displays the PubMed search interface. At the top left is the NCBI logo. In the center is the PubMed logo, and on the right is the National Library of Medicine (NLM) logo. Below the logos is a navigation bar with tabs for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, and Taxonomy. The search bar contains the text "Search PubMed" and a dropdown menu set to "PubMed". The search term "cancer" is entered in the search box, followed by "for". To the right of the search box are "Go" and "Clear" buttons. Below the search box are links for "Limits", "Preview/Index", "History", "Clipboard", and "Details". On the left side, there is a vertical menu with links for "About Entrez", "Text Version", "Entrez PubMed", "Overview", "Help | FAQ", "Tutorial", "New/Noteworthy", "E-Utilities", "PubMed Services", "Journals Database", "MeSH Browser", and "Single Citation". The main content area shows the "PubMed Query:" section with the query text: `("neoplasms"[MeSH Terms] OR cancer[Text Word])`. At the bottom of the query area are "Search" and "URL" buttons.

Geração automática de tesouro

- Tentar gerar um tesouro automaticamente ao analisar a distribuição das palavras em documentos

Geração automática de tesouro

- Tentar gerar um tesouro automaticamente ao analisar a distribuição das palavras em documentos
- Noção fundamental : similaridade entre duas palavras

Geração automática de tesauro

- Tentar gerar um tesauro automaticamente ao analisar a distribuição das palavras em documentos
- Noção fundamental : similaridade entre duas palavras
- Opção 1: duas palavras são **similar se elas co-ocorrem com similar palavras**.

Geração automática de tesauro

- Tentar gerar um tesauro automaticamente ao analisar a distribuição das palavras em documentos
- Noção fundamental : similaridade entre duas palavras
- Opção 1: duas palavras são **similar se elas co-ocorrem com similar palavras**.
 - “carro” \approx “motocicleta ” porque ambos ocorrem com “estrada”, “gasolina” e “carteira”, então eles devem ser similares

Geração automática de tesauro

- Tentar gerar um tesauro automaticamente ao analisar a distribuição das palavras em documentos
- Noção fundamental : similaridade entre duas palavras
- Opção 1: duas palavras são **similar se elas co-ocorrem com similar palavras**.
 - “carro” \approx “motocicleta ” porque ambos ocorrem com “estrada”, “gasolina” e “carteira”, então eles devem ser similares
- Opção 2: duas palavras são **similares se ela ocorrem em uma dada relação gramatical com as mesmas palavras**.

Geração automática de tesauro

- Tentar gerar um tesauro automaticamente ao analisar a distribuição das palavras em documentos
- Noção fundamental : similaridade entre duas palavras
- Opção 1: duas palavras são **similar se elas co-ocorrem com similar palavras**.
 - “carro” \approx “motocicleta ” porque ambos ocorrem com “estrada”, “gasolina” e “carteira”, então eles devem ser similares
- Opção 2: duas palavras são **similares se ela ocorrem em uma dada relação gramatical com as mesmas palavras**.
 - Você pode colher, descascar, comer, cortar, amassar **maçãs** e **peras**

Geração automática de tesouro

- Tentar gerar um tesouro automaticamente ao analisar a distribuição das palavras em documentos
- Noção fundamental : similaridade entre duas palavras
- Opção 1: duas palavras são **similar se elas co-ocorrem com similar palavras**.
 - “carro” \approx “motocicleta ” porque ambos ocorrem com “estrada”, “gasolina” e “carteira”, então eles devem ser similares
- Opção 2: duas palavras são **similares se ela ocorrem em uma dada relação gramatical com as mesmas palavras**.
 - Você pode colher, descascar, comer, cortar, amassar **maçãs** e **peras**
 - ... então **maçãs** e **peras** devem ser similares.

Tesouro baseado em co-ocorrência: exemplos

Palavras	Vizinhos mais próximos
totalmente	completamente, irremediavelmente
mediação	reconciliação, negociação, diplomacia
litografias	desenhos, Picasso, Dali, Gauguin
patógenos	toxinas, bactérias, organismos, parasitas
alface	hortaliças, acelga, hortalã
nada	tudo

Expansão de consultas em motores de busca

- Fonte principal de expansão de consultas em motores de busca: históricos de consultas

Expansão de consultas em motores de busca

- Fonte principal de expansão de consultas em motores de busca: históricos de consultas
- Exemplo 1: Após fazer a consulta [ervas], usuários frequentemente buscam por [ervas medicinais].

Expansão de consultas em motores de busca

- Fonte principal de expansão de consultas em motores de busca: históricos de consultas
- Exemplo 1: Após fazer a consulta [ervas], usuários frequentemente buscam por [ervas medicinais].
 - → “ervas medicinais” é uma potencial expansão de “erva”.

Expansão de consultas em motores de busca

- Fonte principal de expansão de consultas em motores de busca: históricos de consultas
- Exemplo 1: Após fazer a consulta [ervas], usuários frequentemente buscam por [ervas medicinais].
 - → “ervas medicinais” é uma potencial expansão de “erva”.
- Exemplo 2: usuários procurando por [foto de flor] frequentemente clicam no URL photobucket.com/flower. usuários procurando por [desenho de flor] frequentemente clicam no [mesmo URL](#).

Expansão de consultas em motores de busca

- Fonte principal de expansão de consultas em motores de busca: históricos de consultas
- Exemplo 1: Após fazer a consulta [ervas], usuários frequentemente buscam por [ervas medicinais].
 - → “ervas medicinais” é uma potencial expansão de “erva”.
- Exemplo 2: usuários procurando por [foto de flor] frequentemente clicam no URL photobucket.com/flower. usuários procurando por [desenho de flor] frequentemente clicam no [mesmo URL](#).
 - → “desenho de flor” e “foto de flor” são potenciais expansões entre si

Expansão de consultas em motores de busca

- Fonte principal de expansão de consultas em motores de busca: históricos de consultas
- Exemplo 1: Após fazer a consulta [ervas], usuários frequentemente buscam por [ervas medicinais].
 - → “ervas medicinais” é uma potencial expansão de “erva”.
- Exemplo 2: usuários procurando por [foto de flor] frequentemente clicam no URL photobucket.com/flower. usuários procurando por [desenho de flor] frequentemente clicam no [mesmo URL](#).
 - → “desenho de flor” e “foto de flor” são potenciais expansões entre si
 - **Mineração de sequências de cliques.**

Resumo

- Objetivo: aumentar taxa de recuperação
- Opção “local”: retorno de relevância fornecida pelo usuário
 - Com informação sobre documentos, usar algoritmo de Rocchio
- Opção “global”: expansão de consultas com uso de Tesouro
 - Construção com conhecimento do domínio
 - Construção automática
 - Co-ocorrência
 - Análise gramatical
 - Mineração de histórico de consultas